Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Performance and Usability of Code-Free Deep Learning for Chest Radiograph Classification, Object Detection, and Segmentation

Samantha M. Santomartino, BA • Nima Hafezi-Nejad, MD • Vishwa S. Parekh, PhD • Paul H. Yi, MD

From the University of Maryland Medical Intelligent Imaging (UM2ii) Center, Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, 670 W Baltimore St, First Floor, Room 1172, Baltimore, MD 21201 (S.M.S., P.H.Y.); The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, Md (N.H.N., V.S.P.); Department of Computer Science, Whiting School of Engineering (V.S.P.), and Malone Center for Engineering in Healthcare (P.H.Y.), Johns Hopkins University, Baltimore, Md. Received March 28, 2022; revision requested May 11; revision received January 15, 2023; accepted January 26. **Address correspondence to** P.H.Y. (email: *pyi@som.umaryland.edu*).

Supported by Amazon Web Services, proof-of-concept credit granted for investigation of Amazon platform. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2023; 5(2):e220062 • https://doi.org/10.1148/ryai.220062 • Content codes: AI CH

Purpose: To evaluate the performance and usability of code-free deep learning (CFDL) platforms in creating DL models for disease classification, object detection, and segmentation on chest radiographs.

Materials and Methods: Six CFDL platforms were evaluated in this retrospective study (September 2021). Single- and multilabel classifiers were trained for thoracic pathologic conditions using Guangzhou pediatric and NIH-CXR14 (ie, National Institutes of Health ChestX-ray14) datasets, and external testing was performed using subsets of NIH-CXR14 and Stanford CheXpert datasets, respectively. Pneumonia detection and pneumothorax segmentation models were trained using the Radiological Society of North America (RSNA) Pneumonia and Society for Imaging Informatics in Medicine (SIIM) Pneumothorax datasets, respectively. Model performance was evaluated using F1 scores. Usability was evaluated based on feasibility of image uploading and model training, ease of use, and cost.

Results: NIH-CXR14 and CheXpert datasets contained 112 120 (mean age, 47 years ± 17 [SD]; 63 340 male patients) and 151 522 images (mean age, 61 years ± 18; 88 931 male patients), respectively. The other datasets did not report demographics (Guangzhou, 5826 images; RSNA, 26 683 images; SIIM, 15 301 images). Six platforms offered single-label classifiers, four multilabel classifiers, five object detection models, and one segmentation model. Guangzhou pneumonia classifiers demonstrated good internal (F1, 0.93–0.99) and poor external (F1, 0.39–0.44) performance. Multilabel NIH-CXR14 classifiers showed poor internal and external performance (F1, 0.00–0.36 and 0.00–0.76, respectively). NIH-CXR14 single-label classifiers performed poorly (F1, 0.00, all). The single successfully trained pneumonia detection model had an F1 score of 0.48. No segmentation model was successfully trained. Platform usability was limited, with all requiring some type of coded solution.

Conclusion: CFDL platforms demonstrated limited performance and usability for chest radiograph analysis.

Supplemental material is available for this article.

Chest radiographs are the world's most common medical imaging test, with approximately 110 000 000 chest radiographs performed yearly in the United States alone (1). Due to their low cost and usability in a range of disease diagnoses (2,3) and traumatic injuries (4–6), chest radiographs are used globally for both triage and diagnosis of thoracic diseases. However, radiologists available for interpretation are scarce, both in underserved areas of high-income countries, such as rural parts of the United States (7), and in low- or middle-income countries where few radiologists practice (8). Deep learning (DL) systems for automated interpretation and diagnosis of disease on chest radiographs have been proposed as a solution to bring expert-level care to underserved regions (9) while also aiding triage of actionable images (10,11) and augmenting radiologist performance (12–15) in high-service regions.

Although DL has demonstrated expert-level ability to diagnose diseases on medical images (16), including several thoracic conditions on chest radiographs (17–20), DL model development remains inaccessible to those without the coding expertise needed for model training (21,22), including most radiologists and clinicians; thus, DL may be underutilized in health care. Automated machine learning and more specifically, code-free DL (CFDL), seeks to solve this problem by automatically performing DL model training (including decisions traditionally made by a human coder, such as algorithm structure and hyperparameters [23]), in packaged solutions that theoretically bring the power of DL to lay users.

CFDL automated machine learning platforms have shown promise for ophthalmology medical imaging (21,22). However, to our knowledge, similar evaluations for chest radiographs beyond Google's platform or for tasks beyond image classification, such as object detection and segmentation, have not been performed. Furthermore, platform usability has not been extensively evaluated, an important consideration given the theoretical purpose of these platforms to make DL accessible to users without coding expertise.

Abbreviations

AUPRC = area under the precision-recall curve, CFDL = code-free DL, DL = deep learning, GUI = graphic user interface, RSNA = Radiological Society of North America, SIIM = Society for Imaging Informatics in Medicine

Summary

While code-free deep learning platforms performed well in binary classification, they require further development in external testing, usability, and other deep learning tasks before they can be fully implemented in a clinical setting.

Key Points

- Code-free deep learning (CFDL) platforms demonstrated high performance (mean F1, 0.96) on single-label, binary classification tasks for smaller, balanced chest radiograph datasets but did not perform well for more complex datasets and tasks (multilabel classification [F1 score, 0.12], object detection [F1 score, 0.48], and segmentation [not applicable]).
- CFDL usability was greater for image classifiers compared with other tasks, but platforms demonstrated practical feasibility limitations related to data labeling and uploading, model training, and external performance, with all requiring coded solutions.
- Despite great potential for some medical applications, these CFDL platforms are not yet suitable for chest radiograph diagnosis and may have limited accessibility to clinicians without coding experience.

Keywords

Artificial Intelligence, Automated Machine Learning, Chest Radiographs, Deep Learning, Code-Free Deep Learning, Pneumonia, Pneumothorax, Radiology

In this study, we evaluated the performance and usability of six CFDL platforms for development of DL classification, object detection, and segmentation models using publicly available chest radiograph datasets. We report CFDL model performance for diagnosis of disease in adult and pediatric chest radiographs, including external testing results for classifiers. We also report on the usability of the CFDL platforms in terms of feasibility of model development, ease of use, and cost. Given prior works showing that conventional DL models perform variably on external datasets (24), as well as variable performance of Google's AutoML for chest radiograph diagnostic tasks (21), we hypothesized that automated machine learning for chest radiographs would show inconsistent performance between the six CFDL platforms and decreased external performance.

Materials and Methods

Amazon Web Services granted a proof-of-concept credit for investigation of the Amazon platform only; no other funds were provided by industry for this study. The study authors had control of all data and information submitted for publication.

Datasets

In a retrospective study, we chose three DL chest radiograph tasks representing common thoracic diseases with well-established public datasets (Table 1), disease-specific annotations, and prior benchmarks of performance: *(a)* image classification of thoracic diseases (17,18), *(b)* object detection of pneumonia (19), and *(c)* segmentation of pneumothorax (25).

We used 5826 pediatric chest radiographs (Guangzhou pediatric chest radiograph dataset; hereafter, Guangzhou dataset) (26) to develop single-label, binary image classifiers for presence or absence of pneumonia, consistent with previous studies using CFDL platforms (21) and conventional DL (21,27). We used 112 120 chest radiographs (National Institutes of Health ChestX-ray14 dataset; hereafter, NIH-CXR14 dataset) (28) to train multilabel image classifiers (15 labels) on the CFDL platforms offering multilabel classification models (Amazon, Clarifai, Google, Microsoft), consistent with previous studies using CFDL platforms (21) and conventional DL (17). For the two CFDL platforms that did not offer multilabel classification models (Apple and MedicMind), we recategorized the dataset into binary image classifiers (label distribution in Table S8). When applicable, external testing of classification models was performed using 383 pediatric images from the NIH-CXR14 (for Guangzhou models) and 151 522 images from the Stanford CheXpert (29) (for NIH-CXR14 models) datasets (Table 1).

We developed object detection models using 26 683 chest radiographs annotated with pneumonia bounding boxes (Radiological Society of North America [RSNA] Pneumonia Detection Challenge dataset; hereafter, the RSNA dataset) (30). Finally, we developed segmentation models (a pixel-level annotation process) using 15 301 chest radiograph images annotated with pneumothorax segmentations (Society for Imaging Informatics in Medicine [SIIM]/American College of Radiology Pneumothorax Segmentation dataset; hereafter, the SIIM dataset) (31).

Because all images were de-identified and from public, open access databases, no institutional review board approval was required, and the data are Health Insurance Portability and Accountability Act compliant. Additional dataset details are presented in Appendix S1.

CFDL Platform Selection and Model Details

We evaluated six CFDL platforms that were previously identified by Korot et al when evaluating appropriateness for medical imaging (22): Amazon Rekognition Custom Labels, Apple Create ML, Clarifai Train, Google Cloud AutoML Vision, MedicMind DL Training Platform, and Microsoft Azure Custom Vision (Table 2). We note that Korot et al evaluated nonradiologic medical images; additionally, our platform evaluation occurred in September 2021, more than 1 year following that of the previous study, resulting in some feature differences (22) (Table 2). Further system requirements are presented in Appendix S2. Prior to evaluation, a researcher (S.M.S.) with a computer science undergraduate degree spent a minimum of 1.5 hours on each platform exploring and reading provided documentation to understand how each platform worked.

Data were properly labeled according to platform-specific requirements, with multiple options available for several of the platforms (Table S7). We randomly split our data into 80% training, 10% validation, and 10% testing sets, ensuring an equal distribution of labels when manual designation was allowable by platform. When applicable, we

Table 1: Datasets Used to Evaluate Automated Machine Learning Code-Free Deep Learning Platforms

Name	Type	No. of Images	Sex	Mean Age ± SD (y)	Classes
Datasets for internal testing					
Guangzhou pediatric CXR (Guangzhou dataset)	Frontal CXR	5856	NR	NR	Pneumonia (*n* = 4245) Normal (*n* = 1582)
National Institutes of Health (NIH) CXR (NIH-CXR14 dataset)	Frontal CXR	112 120	63 340 male patients	47 ± 17	Atelectasis (*n* = 11 559) Cardiomegaly (*n* = 2776) Effusion (*n* = 13 317) Infiltration (*n* = 19 894) Mass (*n* = 5782) Nodule (*n* = 6331) Pneumonia (*n* = 1431) Pneumothorax (*n* = 5302) Consolidation (*n* = 4667) Edema (*n* = 2303) Emphysema (*n* = 2516) Fibrosis (*n* = 1686) Pleural thickening (*n* = 3385) Hernia (*n* = 227) No findings (*n* = 60 361)
Radiological Society of North America (RSNA) Pneumonia Detection Challenge (RSNA dataset)	Frontal CXR	26 683	NR	NR	Pneumonia (*n* = 6012) No pneumonia (*n* = 20 671)*
SIIM-ACR pneumothorax segmentation (SIIM dataset)	Frontal CXR	15 301	NR	NR	Pneumothorax (*n* = 3577) Not pneumothorax (*n* = 9378)
Datasets for external testing					
NIH-CXR14†	Frontal CXR	383	202 male patients	3 ± 2	Pneumonia (*n* = 107) Normal (*n* = 276)
CheXpert‡	Frontal CXR	151 522	88 931 male patients	61 ± 18	Atelectasis (*n* = 29 795) Cardiomegaly (*n* = 23 451) Effusion (*n* = 76 963) Pneumonia (*n* = 4683) Pneumothorax (*n* = 17 700) Consolidation (*n* = 13 015) Edema (*n* = 49 717) No findings (*n* = 17 000)

Note.—ACR = American College of Radiology, CXR = chest radiograph, NA = not applicable, NR = not reported, SIIM = Society for Imaging Informatics in Medicine.
* The negative images were discarded from object detection model training because the code-free deep learning platforms did not allow for training with negative images (ie, those without a positive bounding box).
† Only a subset of NIH-CXR14 was used (pediatric patients aged 1 to 5 years with pneumonia or normal readings).
‡ Only a subset of CheXpert was used (thoracic diagnosis labels that were common with NIH-CXR14).

also preserved the splits established by the dataset publishers. All six platforms offered a graphic user interface (GUI) for data upload, as well as various other options including command line interface prompts and client library calls. We used several different methods for data upload into the corresponding CFDL platform (see Results). Additional dataset and label processing and splitting details are presented in Appendix S2.

One model was trained per dataset-platform-model triad ("model" refers to either classification, object detection, or segmentation). All automated machine learning models were trained for the maximum number of hours allowed within each platform's free tier or that not exceeding $100 (22). Platforms performed early stopping if model performance plateaued. Specific details about the configuration of model training time limits are presented in Appendix S2.

Evaluation of CFDL Performance

F1 was the only performance metric common to all platforms, provided directly (Amazon, Google) or calculable from data outputs (Apple, Clarifai, Microsoft, MedicMind). Models allowing for threshold selection (Clarifai, Google, Microsoft) were evaluated at the default value (0.5) (22). F1 scores were summarized and compared across all platforms using the arith-

Table 2: Code-Free Deep Learning Platform Features

Feature	Amazon	Apple	Clarifai	Google	MedicMind	Microsoft
Classification (C), multilabel classification (MC), object detection (OD), segmentation (S)	C, MC, OD	C*, OD	C, MC	C, MC, OD	C*, OD, S	C, MC, OD
Supported image types	jpeg, png	Any format compatible with Quicktime Player (jpeg, png)	jpeg, png, tiff, bmp, webp	jpeg, png, gif, bmp†, ico†	jpg, png, ppm, tiff, zip, gif, dcm, csv, txt	jpg, png, bmp, gif
CSV image label upload	N	N	N	Y	N*	N
Cloud bucket for image storage	Y	N	N	Y	N	N
Manual train and/or test split	Y	Y	N	Y	N*	N
Manual validation set designation	N	Y*	N	Y	N	N
Designation of training hours	N	N	N	Y	N	Y
Batch prediction (external test)	N	N	N	Y	Y	N
Generates confusion matrix	Y*‡	N*	Y	Y	Y*‡	N
Evaluation metrics (other)	Precision, recall, F1, assumed threshold	Precision, recall, accuracy	AUC, precision, recall	AUPRC, precision, recall, F1, confidence thresholds and/or curves	Accuracy, recall, specificity	AUPRC, precision, recall, F1
Live adjustable prediction thresholds	N	N	Y	Y	N	Y
Ability to download model	N§	Y*	N	Y	N§	Y
Free tier quotas	10 training hours and 4 prediction hours a month for 3 months	NA	1000 operations and 10 000 input objects per month	40 free node hours each for training and online prediction, and 1 free node hour for batch prediction	NA	Training: 1 hour per month 5000 images per project Online prediction: 10 000 predictions per month

Note.—AUC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve, CSV = comma separated values file, N = no, NA = not applicable, Y = yes.
* Indicates updates and/or differences between our findings and those reported in the table by Korot et al (22).
† Image format supported for training only, not predictions.
‡ Confusion matrix not directly given but can be calculated by other given output.
§ Does not allow model downloading but does offer a method for model deployment.

metic mean with 95% CI. As the mean of precision and recall, F1 is a widely used and useful statistic for comparing the performance of DL platforms (32). As previous work evaluating the Google CFDL platform for chest radiograph classification using both Guangzhou and NIH-CXR14 datasets reported area under the precision recall-curve (AUPRC) (21), we report AUPRC when possible to allow for comparisons. Other metrics including accuracy, positive predictive value, and negative predictive value are reported when provided by the platform.

Evaluation of CFDL Usability

To assess usability, we first evaluated feasibility, defined as the ability of a given CFDL platform to easily upload data and annotations and to train a model for a target task. If data and/or annotation uploading was not feasible using a codeless solution, we used code to evaluate for model training feasibility separately from the data and/or annotation uploading portions. In the event of the platform crashing or otherwise failing, we attempted to train the model at least three times before

concluding it was not feasible. We next evaluated the ability of CFDL platforms to perform external testing on trained models, an important feature for clinical generalizability (33,34). Finally, we evaluated the costs of each CFDL platform for these diagnostic tasks. In the event of technical issues, we consulted each company's technical support to ensure that we had made every effort to properly use the platforms.

Statistical Analysis

Raw model image-level prediction outputs were not provided by or accessible from any of the CFDL platforms, precluding more granular statistical evaluation. Statistical calculations were used solely to manipulate the platform output into the desired metric, utilizing Google Sheets (version 96.0.4664.55) and R software (version 4.1.1; R Foundation for Statistical Computing). All statistics were reviewed by a statistical consultant (N.H.N.).

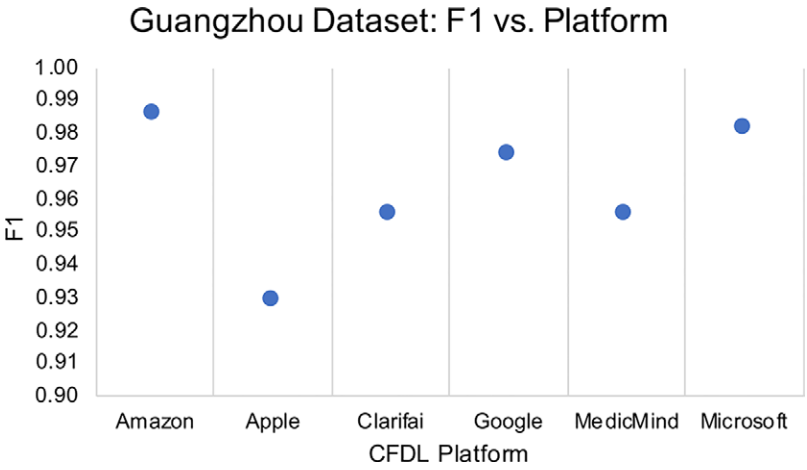

Figure 1: Graph shows F1 scores for single-label classification of pediatric pneumonia on the Guangzhou dataset per code-free deep learning (CFDL) platform.

Results

Dataset Demographics

Two datasets reported demographics: NIH (mean age, 47 years ± 17 [SD]; 63 340 male patients) and CheXpert (mean age, 61 years ± 18; 88 931 male patients). The other datasets did not report demographics, although Guangzhou images were all from patients aged 1–5 years (26) (Table 1).

Model Performance

Guangzhou dataset: Single-label classification.— F1 scores for single-label classification of pediatric pneumonia for the Guangzhou dataset were uniformly high across all platforms with a mean F1 score of 0.96 (95% CI: 0.94, 0.99): Amazon, 0.99; Apple, 0.93; Clarifai, 0.96; Google, 0.97; MedicMind, 0.96; and Microsoft, 0.98 (Fig 1). Accuracy was also high with the average across platforms achieving 94.6%. Two platforms reported AUPRC: Google, 0.99, and Microsoft, 0.98. Remaining performance metrics are summarized in Table 3 and are compared with prior automated machine learning and conventional DL literature in Table S1.

External testing on supporting platforms (Google, MedicMind) showed poor generalizability to a pediatric subset of the NIH-CXR14 dataset, as F1 scores were less than 0.5 (mean F1, 0.41; 95% CI: 0.093, 0.73), representing decrease in performance greater than 0.4 (Table S2).

NIH-CXR14 dataset: Multilabel and single-label classification.— Multilabel classification models trained on the NIH-CXR14 dataset showed uniformly poor performance, with a mean F1 score of 0.12 (95% CI: -0.14, 0.38). Performance re-

Table 3: Summary of Single-Label, Binary Classification Performance of Algorithms Trained on the Guangzhou Dataset for All Six Code-Free Deep Learning Platforms

Platform	TP	FP	TN	FN	PPV (%)	Sensitivity (%)	Specificity (%)	NPV (%)	Accuracy (%)	F1
Amazon	383	4	230	7	99.0 (383/387)	98.2 (383/390)	98.3 (230/234)	97.0 (230/237)	98.2 (612/624)	0.996
Apple	NR	NR	NR	NR	90.0 (NR/NR)	96.0 (NR/NR)	NR (NR/NR)	NR (NR/NR)	NR (NR/NR)	0.93
Clarifai	698	29	237	37	96.0 (698/727)	95.0 (698/735)	89.1 (237/266)	86.5 (237/274)	93.4% (935/1001)	0.96
Google	423	22	212	1	95.1 (423/445)	99.8 (423/424)	90.6 (212/234)	99.5 (212/213)	96. (635/658)	0.97
MedicMind	791	2	319	72	99.7 (791/793)	91.7 (791/863)	99.4 (319/321)	81.6 (319/391)	93.8 (1110/1184)	0.96
Microsoft	NR	NR	NR	NR	97.8 (NR/NR)	98.5 (NR/NR)	NR (NR/NR)	NR (NR/NR)	NR (NR/NR)	0.98

Note.—Unless otherwise indicated, data are numbers or percentages with numerators and denominators in parentheses. The Guangzhou dataset was used to train for classification of pneumonia or no pneumonia. Microsoft and Apple do not provide image level results, precluding calculation of accuracy, specificity, and NPV. FN = false negative, FP = false positive, NPV = negative predictive value, NR = not reported, PPV = positive predictive value, TN = true negative, TP = true positive.

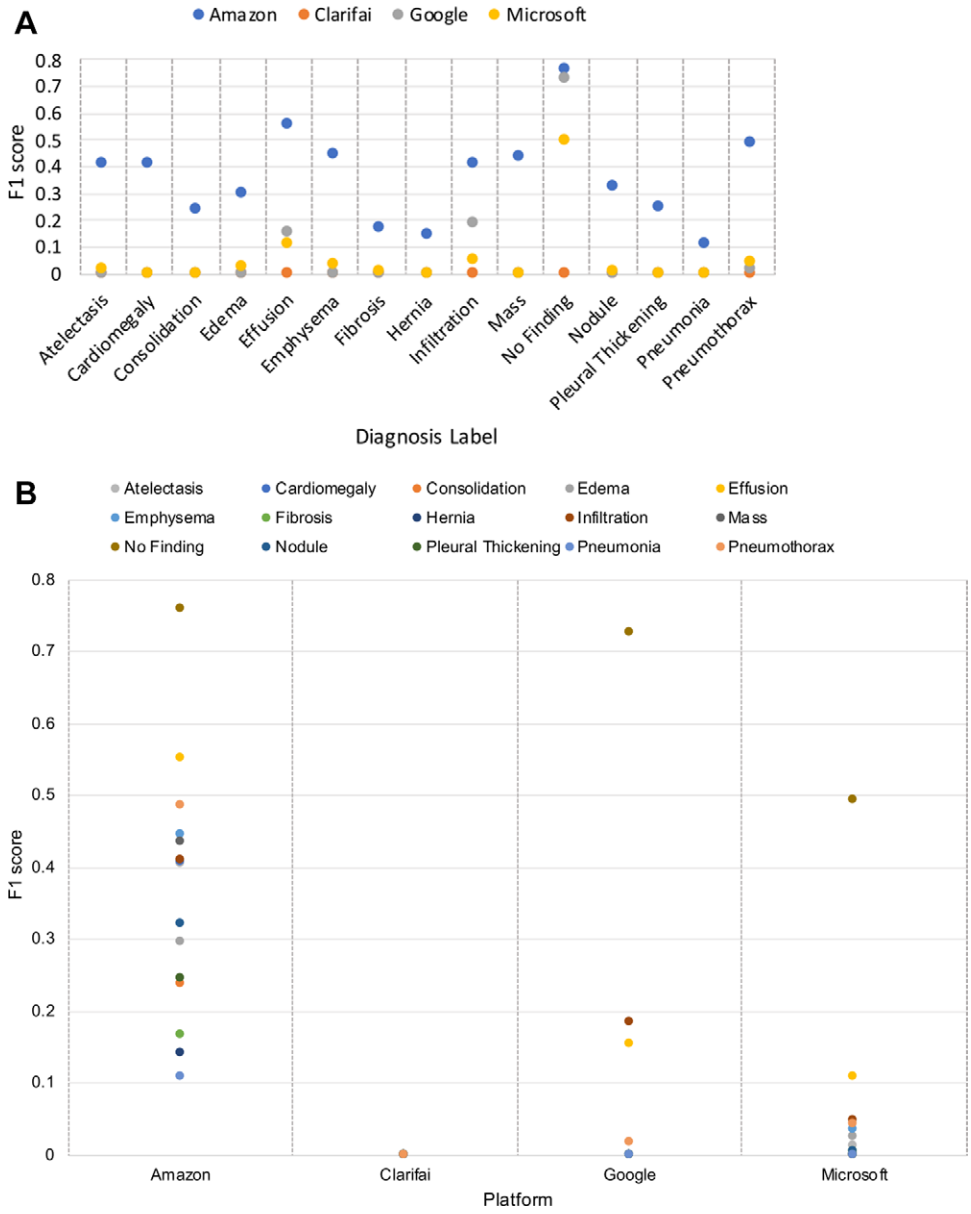

Figure 2: Graphs show multilabel image classification performance of algorithms trained on the NIH-CXR14 dataset. F1 scores grouped by **(A)** diagnosis label and **(B)** code-free deep learning platform. CXR = chest radiograph, NIH = National Institutes of Health.

ported as mean F1 (mean precision, mean recall) for each platform was as follows: Amazon, 0.36 (31.81%, 43.30%); Clarifai, 0.00 (0.00%, 0.00%); Google, 0.072 (7.45%, 8.63%); and Microsoft, 0.052 (30.51%, 3.41%). Amazon consistently had higher F1 scores in the multilabel image classification model, although limited model output precluded statistical comparisons (Fig 2). When omitting the label of "no finding" and only considering the explicit disease labels, mean F1 performance dropped further (0.095; 95% CI: -0.16, 0.35). AUPRC for platforms reporting it demonstrated poor performance (Google, 0.52; Microsoft, 0.48). Although F1 scores for specific labels were low, negative predictive value was high for diseases (>70%), and F1 scores for no finding labels were greater than 0.70 for Amazon and Google (Fig 2). Detailed multilabel classification results are summarized in Table S3A–S3D.

External testing of multilabel image classification performed using the Stanford CheXpert (29) dataset on the sole supporting platform (Google) was limited as the mean F1 was 0.059 (95% CI: -0.049, 0.17; mean precision, 13.6%; mean recall, 7.11%) (Table S4).

For single-label classification of the NIH-CXR14 dataset (performed on the two platforms [Apple, MedicMind] without multilabel classification), only two MedicMind models (atelectasis and cardiomegaly) were successfully trained due to GUI crashes. These models performed poorly, with F1 scores of 0 for both (Table S5). External testing of these models also failed due to repeated GUI crashes.

RSNA dataset: Object detection.— Although Amazon, Apple, Google, and Microsoft offered object detection, only

Table 4: Successful Models Trained by Task and Platform

Platform	Single-Label Classification (Guangzhou)	Multilabel or Single-Label* Classification (NIH-CXR14)	Object Detection (RSNA)	Segmentation (SIIM)
Amazon	S	S	S⁺	NA
Apple	S	NS	NS	NA
Clarifai	S	S	NA	NA
Google	S	S	S	NA
MedicMind	S	Partial (2/15 possible models trained successfully)	NS	NS
Microsoft	S	S	NS	NA

Note.—CXR = chest radiograph, NA = not applicable, NIH = National Institutes of Health, NS = not successfully completed, RSNA = Radiological Society of North America, S = successfully completed, SIIM = Society for Imaging Informatics in Medicine, S⁺ = successfully completed but not clinically relevant or usable.
* Multilabel classifiers were offered by four platforms (Amazon, Clarifai, Google, and Microsoft). The remaining two platforms (Apple and MedicMind) only offered single-label classifiers.

the Google platform was successful in creating a clinically relevant pneumonia detection model. This model showed overall poor performance, with an F1 score of 0.48 and recall of 36.3%, albeit with relatively high precision of 71.3%, at default confidence and intersection over union thresholds of 0.5 each.

SIIM dataset: Segmentation.— A segmentation model was not successfully trained on the SIIM dataset using the sole CFDL platform offering segmentation models (MedicMind).

Usability of CFDL Platforms

Feasibility: Data labeling.— Reorganizing the datasets, performing label splits, and creating unique annotation files for each model was not possible for any CFDL platform or task without some coding intervention. Microsoft object detection and MedicMind object detection and segmentation tasks were infeasible due to manual labeling requirements. Amazon object detection was infeasible due to the inability to detect multiple instances of a given object class.

Feasibility: Data upload.— All CFDL platforms offered a codeless GUI for data upload. However, for Amazon, Clarifai, Google, and Microsoft, data upload was infeasible using the GUI alone due to prohibitively time-consuming and cumbersome processes with low data size limitations. We therefore used code-based solutions for data upload to these four CFDL platforms. Although coding knowledge was required, these solutions were considerably faster than the time required by the GUIs. MedicMind's GUI worked sufficiently well for ease of use and the speed to upload most datasets. However, it repeatedly crashed and failed to complete upload of the largest dataset (NIH-CXR14: 112 120 images, 42 GB). Unlike the other platforms, MedicMind offers no coding solution for data upload. Only Apple allowed completely codeless data upload, as it was the only locally run platform; no data were uploaded onto a server, only into the application itself.

Feasibility: Model training.— Image classification was overall more successful than object detection and segmentation (Table 4). For the Guangzhou single-label classification of pediatric pneumonia, all six CFDL platforms successfully trained a single-label image classification model. For the NIH-CXR14 multilabel classification, all four CFDL platforms offering this feature successfully trained models. For the two CFDL platforms that offered only single-label classification (Apple and MedicMind), only two models were successfully trained (both MedicMind) due to GUI crashes. Only the Google object detection model was successful in training a clinically relevant solution, and no segmentation model was successfully trained.

External testing.— Google and MedicMind offered external testing, however, both had feasibility issues. MedicMind was unable to process batches of more than 400 images without crashing. Google offered only a code-based solution for external testing of image batches, working for the subset of pediatric NIH-CXR14 images (383 images) used to validate the Guangzhou model but crashing after evaluating 4763 of the 151 521 CheXpert images used to validate the NIH-CXR14 model.

Cost.— The CFDL platforms had both free and paid tier options (Table S6). All models were completed within our cost limit of $100 per model. Two platforms (Apple and MedicMind) were free. All Clarifai models were trained successfully using the free tier. Amazon, Google, and Microsoft had costs of less than $100 to train some models. We used free credits and tiers when possible. Model training was the most expensive task, although image storage, application transactions, and batch predictions were also costly.

An expanded usability evaluation with technical details is provided in Appendix S2.

Discussion

We evaluated performance and usability of six CFDL platforms for development of DL models for chest radiograph classifica-

tion, object detection, and segmentation. Model performance was variable, with single-label pediatric pneumonia classification outperforming multilabel disease classification and object detection models. On external testing, however, both single-label and multilabel classification models performed poorly. No segmentation model was successfully trained. Platform usability was limited, often requiring coded solutions with variable success rates in training models. Although CFDL platforms present a compelling use case, our results indicate that they are not currently suitable for clinical use or easily accessible to clinicians without coding experience.

Guangzhou classification performance was high (mean F1, 0.96; accuracy, 94.6%) across all platforms and comparable to previous Google AutoML (F1, 0.98; accuracy, 97.2%) (21) and conventional DL (F1, 0.96; accuracy, 92.8% [calculated from Kermany et al results]) (18) models. Poor performance of our multilabel NIH-CXR14 models (AUPRC, 0.52) was similar to that of a prior Google AutoML model (AUPRC, 0.57 [21]). When compared with higher-performing Guangzhou models, the relatively larger dataset size for this task suggests that large dataset size is a limiting factor for these platforms. The average F1 score of these multilabel classification models was 0.095 when the *"no findings"* label was omitted. This is comparable to prior conventional DL models that achieved an average F1 score of 0.082 when trained on the same NIH-CXR14 dataset (35), suggesting that multilabel classification is a complex task for DL models trained using both automated machine learning and human-engineered methods. All classifiers performed poorly on external testing, echoing previous findings of limited external generalizability of DL pneumonia classification models trained on these datasets (27,36). Potential reasons for the poor classifier performance include differences in disease prevalence between hospitals (36) and the anatomic regions included in the field of view for images from different sites (27).

To our knowledge, our study is the first assessment of CFDL platforms for nonclassification tasks. Unfortunately, only Google object detection resulted in a clinically useful model. Although this model performed poorly overall, precision was relatively high (71.3%). However, this performance is likely exaggerated, as the CFDL platform did not allow "negative" images for training (77.5% [20 671 of 26 683] of RSNA images were negative). This positive sample-enriched dataset likely accounts for the high precision compared with state-of-the-art results in the RSNA Pneumonia Detection Challenge, where the best model achieved an average precision of 25.5% (19).

CFDL feasibility was greater for image classifiers compared with other tasks but may still be inaccessible to clinicians without coding experience. The platforms frequently required coding for data organization and upload, and repeatedly crashed when uploading larger datasets, experiences consistent with previous works (21,22). Apple and MedicMind single-label NIH-CXR14 classification models crashed repeatedly, similar to experiences by Korot et al using MedicMind models for ophthalmologic images (22). Such infeasibility may be attributed to the relative size of the training dataset and complexity of the task. While binary classifiers for our smaller dataset (Guangzhou) were generally successful, the same models on our larger dataset (NIH-CXR14), and more computationally intensive tasks (object detection, segmentation), most often failed secondary to platform crashes. Accordingly, caution is warranted when using CFDL platforms with large datasets or for nonclassification use cases.

External testing is important to evaluate for clinical generalizability (24,33). Only two platforms (Google and MedicMind) featured external testing, both with limitations; Google required coding, and both platforms crashed. We cannot overstate the importance of external testing, as performance may be overestimated on internal test sets (as we found). Similar to Korot et al (22), we cannot recommend CFDL platforms that do not have an external testing feature.

Successful models were completed at reasonable cost (<$100 per model), similar to previous CFDL studies using the free tier only (21) or with a maximum of $100 per model (22). This supports the feasibility of CFDL platforms from a resource standpoint. We highlight that the CFDL platforms are the same in terms of features between free and paid tiers; the paid tiers begin after initial free credits are used up, with no change to the actual platform.

While no platforms are currently suitable for clinical use, the most promising platforms appear to be Amazon and Google both in usability (they were the most easily used platforms with the greatest number of successfully trained models) and performance, having the highest internal testing F1 scores on binary classification. Furthermore, as both companies are industry leaders in the cloud computing space, the continued iteration and improvement of these platforms is expected, although time will confirm. Based on our findings, key improvements to make these platforms more clinically useful include providing a more user-friendly and stable interface for imaging data and annotation uploads that do not require coding (though these platforms are generally more accessible than conventional DL, as model training and testing requires no coding knowledge) and allowing extended external testing. Technically, key improvements include more granular statistical outputs for model performance (eg, at the image level) to allow end-users to perform their own statistical analyses and explainability tools (eg, saliency maps) to allow for "sanity checks" of model performance.

Our study had limitations. First, our results may not generalize to CFDL platforms beyond those evaluated. However, as these were previously identified by Korot et al after a systematic search (22) and represented four major cloud computing providers, we felt that they represented the most well-developed platforms available. Second, reported performance metrics were variable, limiting clinically meaningful comparisons. Additionally, we were unable to compare the performance of our object detection task to a traditional DL model due to lack of available data and comparable performance metrics for similar models. Third, because training and testing data splits used in previous studies were impossible to replicate (17,18,21,22), the performance results are not a direct comparison (nor is it possible to train exactly comparable conventional DL models, as CFDL data splits are not provided by the platforms). However, we maintained similar split percentages to create comparable results. Finally, CFDL platforms, like all DL algorithms, are "black boxes" compounded by the automated aspect of automated

machine learning, which precludes fine-tuning of standard features or fine-grained interrogation of learning curves to evaluate for adequate model training. We emphasize that these models lack transparency with regard to model training and monitoring and do not report standard details, such as image preprocessing steps and specific models and/or hyperparameters used, which is a major limitation of these platforms. Although Korot et al (22) reported saliency map generation from the MedicMind platform as an explainability tool, the version of MedicMind that we evaluated offered saliency map generation only for disease severity grading classifiers.

Although automated machine learning CFDL platforms promise to democratize DL to noncoding experts, including radiologists, their usability and performance is generally limited, often requiring coding knowledge for practical use and failing to successfully train high-performing models. We thus recommend caution in using these CFDL platforms for medical image analysis and advocate for greater collaboration between artificial intelligence developers and clinicians.

Author contributions: Guarantors of integrity of entire study, **S.M.S., P.H.Y.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, all authors; clinical studies, **P.H.Y.**; experimental studies, **S.M.S., P.H.Y.**; statistical analysis, all authors; and manuscript editing, all authors

Disclosures of conflicts of interest: **S.M.S.** No relevant relationships. **N.H.N.** No relevant relationships. **V.S.P.** No relevant relationships. **P.H.Y.** Associate editor of *Radiology: Artificial Intelligence.*

References

1. Mettler FA Jr, Mahesh M, Bhargavan-Chatfield M, et al. Patient Exposure from Radiologic and Nuclear Medicine Procedures in the United States: Procedure Volume and Effective Dose for the Period 2006-2016. Radiology 2020;295(2):418–427.
2. Wong HYF, Lam HYS, Fong AH-T, et al. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. Radiology 2020;296(2):E72–E78.
3. Toussie D, Voutsinas N, Finkelstein M, et al. Clinical and Chest Radiography Features Determine Patient Outcomes in Young and Middle-aged Adults with COVID-19. Radiology 2020;297(1):E197–E206.
4. Talbot BS, Gange CP Jr, Chaturvedi A, Klionsky N, Hobbs SK, Chaturvedi A. Traumatic Rib Injury: Patterns, Imaging Pitfalls, Complications, and Treatment. RadioGraphics 2017;37(2):628–651.
5. Newbury A, Dorfman JD, Lo HS. Imaging and Management of Thoracic Trauma. Semin Ultrasound CT MR 2018;39(4):347–354.
6. Ho M-L, Gutierrez FR. Chest radiography in thoracic polytrauma. AJR Am J Roentgenol 2009;192(3):599–612.
7. Rosenkrantz AB, Hughes DR, Duszak R Jr. The U.S. Radiologist Workforce: An Analysis of Temporal and Geographic Variation by Using Large National Datasets. Radiology 2016;279(1):175–184.
8. Ali FS, Harrington SG, Kennedy SB, Hussain S. Diagnostic Radiology in Liberia: A Country Report. J Glob Radiol 2015;1(2):6.
9. Mollura DJ, Culp MP, Pollack E, et al. Artificial Intelligence in Low- and Middle-Income Countries: Innovating Global Health Radiology. Radiology 2020;297(3):513–520.
10. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. Radiology 2019;291(1):196–202.
11. Jang SB, Lee SH, Lee DE, et al. Deep-learning algorithms for the interpretation of chest radiographs to aid in the triage of COVID-19 patients: A multicenter retrospective study. PLoS One 2020;15(11):e0242759.
12. Nam JG, Park S, Hwang EJ, et al. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology 2019;290(1):218–228.
13. Kim W, Lee SM, Kim JI, et al. Utility of a Deep Learning Algorithm for Detection of Reticular Opacity on Chest Radiograph in Patients with Interstitial Lung Disease. AJR Am J Roentgenol 2022;218(4):642–650.
14. Yi PH, Kim TK, Yu AC, Bennett B, Eng J, Lin CT. Can AI outperform a junior resident? Comparison of deep neural network to first-year radiology residents for identification of pneumothorax. Emerg Radiol 2020;27(4):367–375.
15. Homayounieh F, Digumarthy S, Ebrahimian S, et al. An Artificial Intelligence-Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. JAMA Netw Open 2021;4(12):e2141096.
16. Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med 2018;24(5):539–540.
17. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15(11):e1002686.
18. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172(5):1122–1131.e9.
19. Pan I, Cadrin-Chênevert A, Cheng PM. Tackling the Radiological Society of North America Pneumonia Detection Challenge. AJR Am J Roentgenol 2019;213(3):568–574.
20. Hurt B, Yen A, Kligerman S, Hsiao A. Augmenting Interpretation of Chest Radiographs With Deep Learning Probability Maps. J Thorac Imaging 2020;35(5):285–293.
21. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health 2019;1(5):e232–e242.
22. Korot E, Guan Z, Ferraz D, et al. Code-free deep learning for multi-modality medical image classification. Nat Mach Intell 2021;3(4):288–298.
23. Janakiram MSV. Why AutoML Is Set To Become The Future Of Artificial Intelligence. Forbes. https://www.forbes.com/sites/janakirammsv/2018/04/15/why-automl-is-set-to-become-the-future-of-artificial-intelligence/. Published April 15, 2018. Accessed July 26, 2021.
24. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14(1):40.
25. Filice RW, Stein A, Wu CC, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. J Digit Imaging 2020;33(2):490–496.
26. Kermany D. Large dataset of labeled optical coherence tomography (OCT) and Chest X-Ray images. Mendeley; 2018.
27. Xin KZ, Li D, Yi PH. Limited generalizability of deep learning algorithm for pediatric pneumonia classification on external data. Emerg Radiol 2022;29(1):107–113.
28. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; 2097–2106. http://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html.
29. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proc AAAI Conf Artif Intell 2019;33(01):590–597.
30. RSNA Pneumonia Detection Challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. Accessed November 19, 2021.
31. SIIM-ACR Pneumothorax Segmentation. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/overview/description. Accessed November 19, 2021.
32. Erickson BJ, Kitamura F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiol Artif Intell 2021;3(3):e200126.
33. Bluemke DA, Moy L, Bredella MA, et al. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers-From the Radiology Editorial Board. Radiology 2020;294(3):487–489.
34. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2020;2(2):e200029.
35. Zech J. reproduce-chexnet: Reproduce CheXNet. Github; https://github.com/jrzech/reproduce-chexnet. Accessed November 17, 2022.
36. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.